Opinion Summarization Using Entity Features and Probabilistic Sentence Coherence Optimization

(UIUC at TAC 2008 Opinion Summarization Pilot)

Nov 19, 2008

Hyun Duk Kim, Dae Hoon Park,
V.G.Vinod Vydiswaran, ChengXiang Zhai

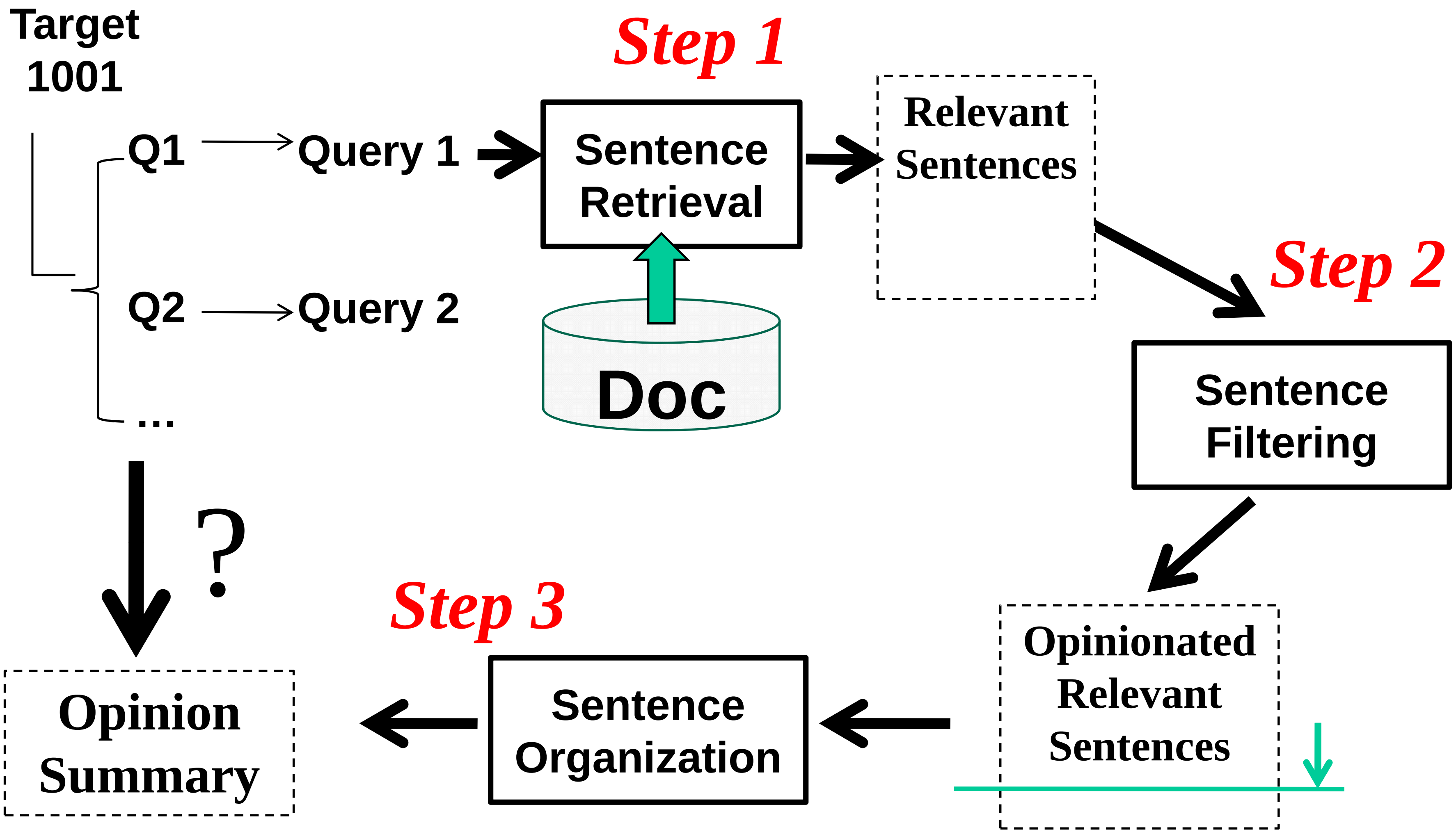
Department of Computer Science

University of Illinois, Urbana-Champaign

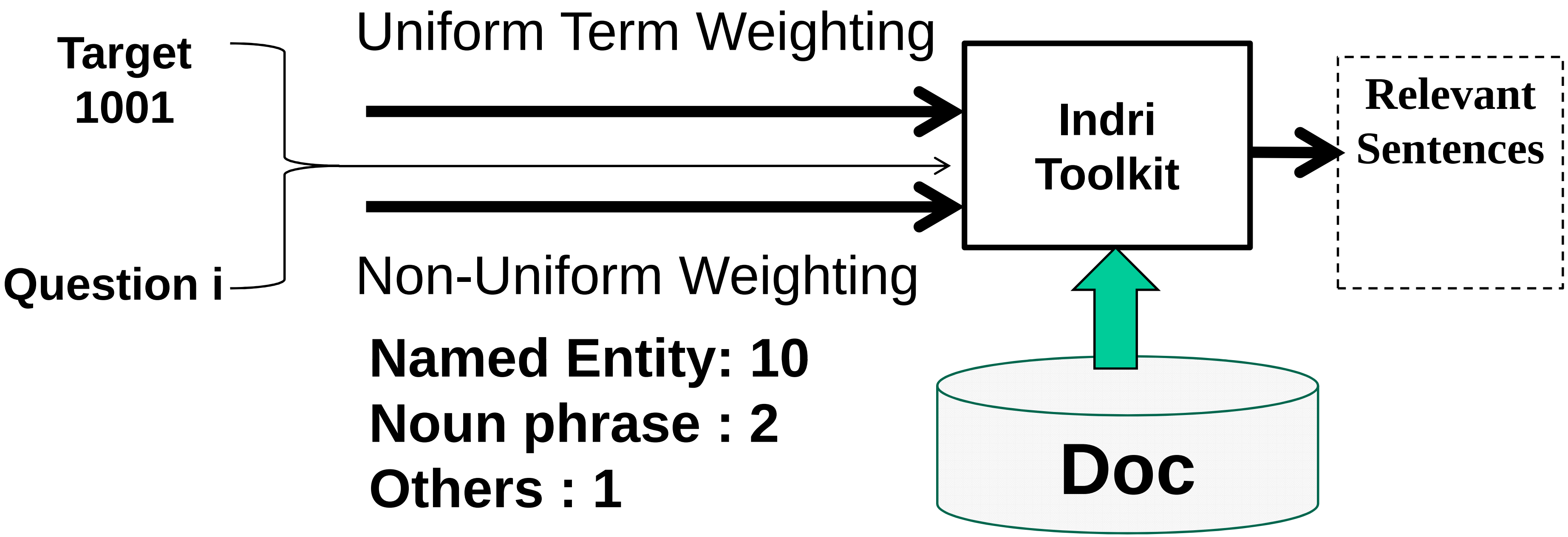
Research Questions

1. Can we improve sentence retrieval by assigning more weights to entity terms?
2. Can we optimize the coherence of a summary using a statistical coherence model?

General Approach



Step 1: Sentence Retrieval



Step 2: Sentence Filtering

Relevant Sentences	Keep only **Opinionated** Sentences	Keep only the **same polarity**	**Remove redundancy**
S1	S1	S1	S1
S2	~~S2~~	~~S2~~	~~S2~~
S3	S3	S3	S3
S1’	S1’	S1’	~~S1’~~
S2’	~~S2’~~	~~S2’~~	~~S2’~~
S4	S4	~~S4~~	~~S4~~
S6	S6	S6	S6
S7	S7	S7	~~S7~~
S4’	S4’	S4’	~~S4’~~

The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign Large Scale Information Management

TIMan

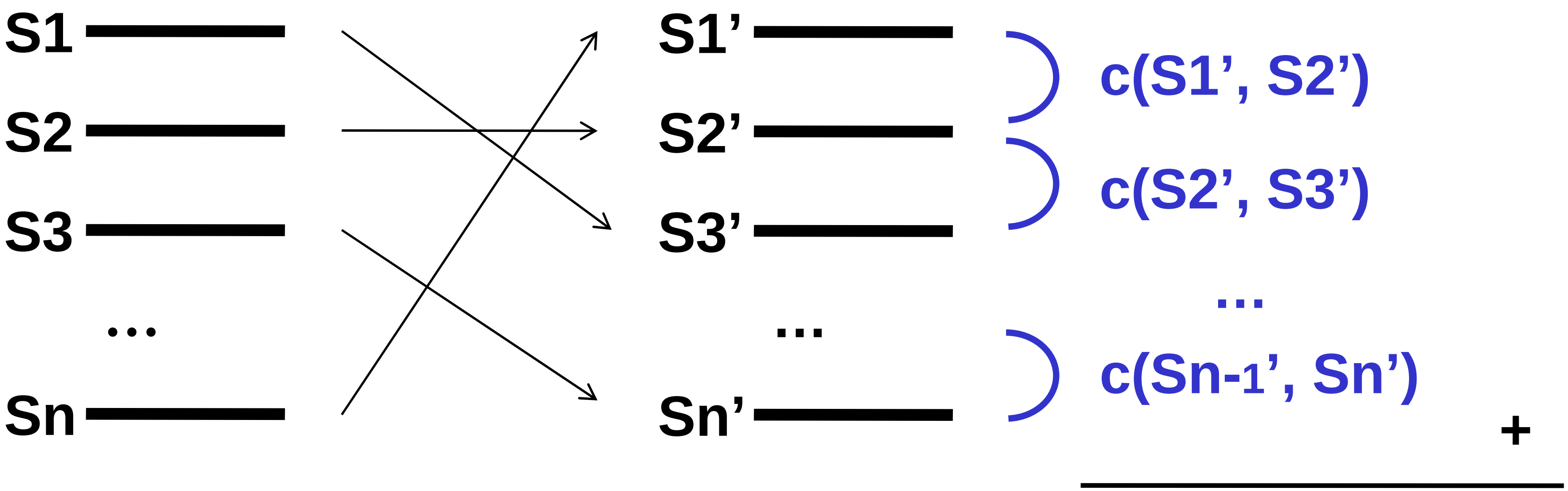
Step 3: Summary Organization
(Method 1: Polarity Ordering)

– **Paragraph structure by question and polarity**
– **Add guiding phrase**

The first question is …
Following are positive opinions…
Following are negative opinions…
The second question is …
Following are mixed opinions…
…

Step 3: Summary Organization

(Method 2: Statistical Coherence Optimization)



- Coherence function: $c(S_i, S_j)$
- Use a greedy algorithm to order sentences to maximize the total score

$$c(S_1', S_2') + c(S_1', S_2') + \ldots + c(S_{n-1}', S_n')$$

Probabilistic Coherence Function

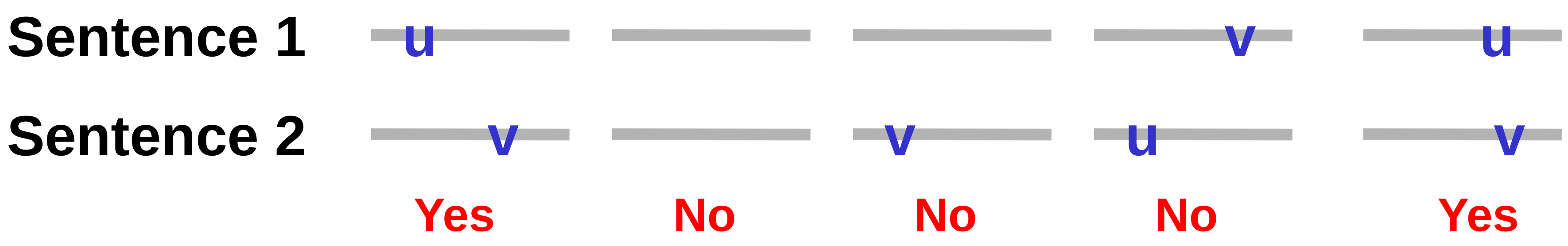
(Idea similar to [Lapata 03])

$$c(s_i, s_j) = \sum_{u \in s_i, v \in s_j} \frac{p(u,v)}{|s_i||s_j|}$$

Average coherence probability (Pointwise mutual information) over all word combinations

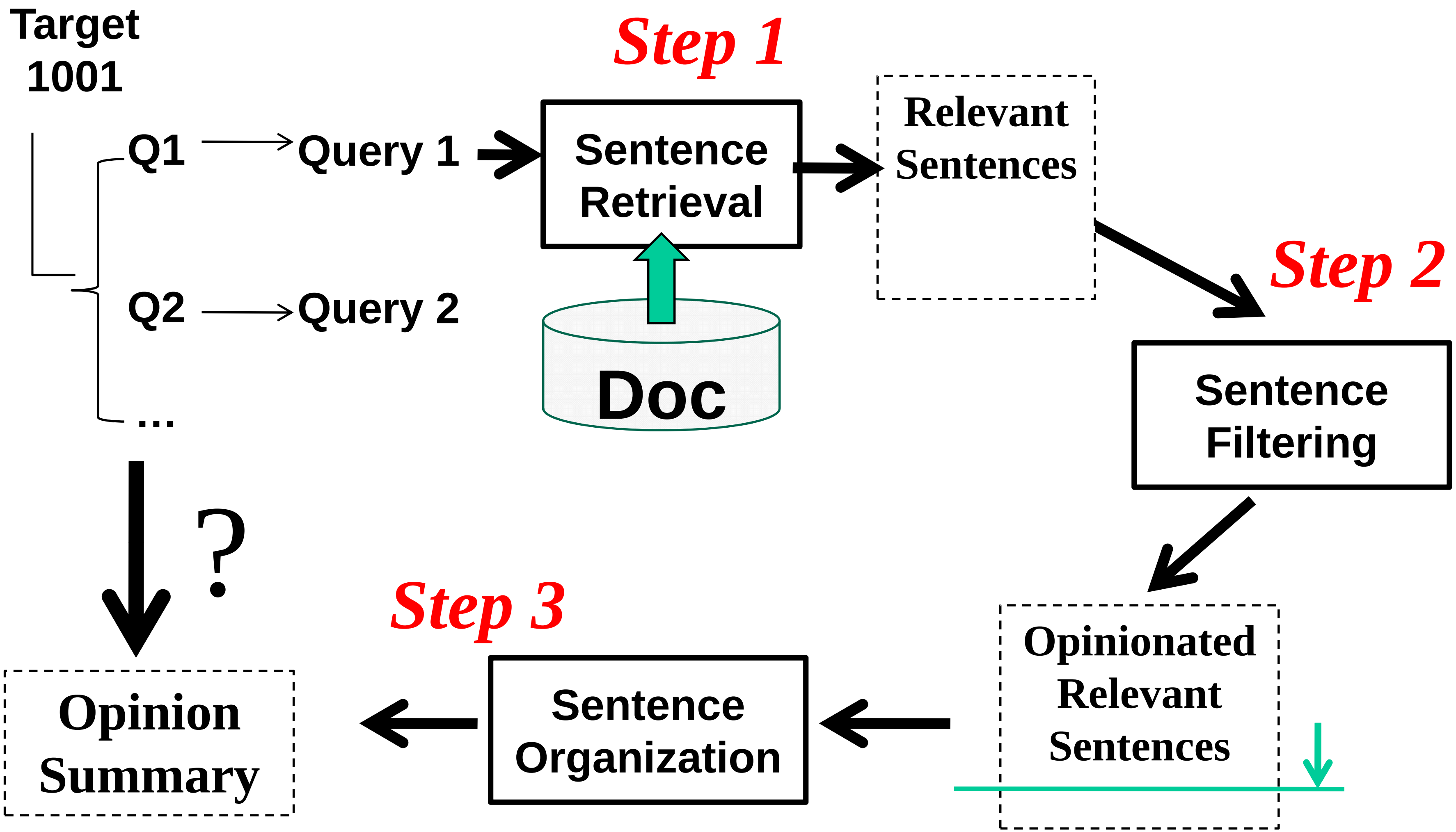
$$\hat{p}(u,v) = \frac{count(\text{"}u \text{ and } v \text{ in two adjacent sentences"}) + 0.001}{count(u) \cdot count(v) + 1.0}$$

Train with original document

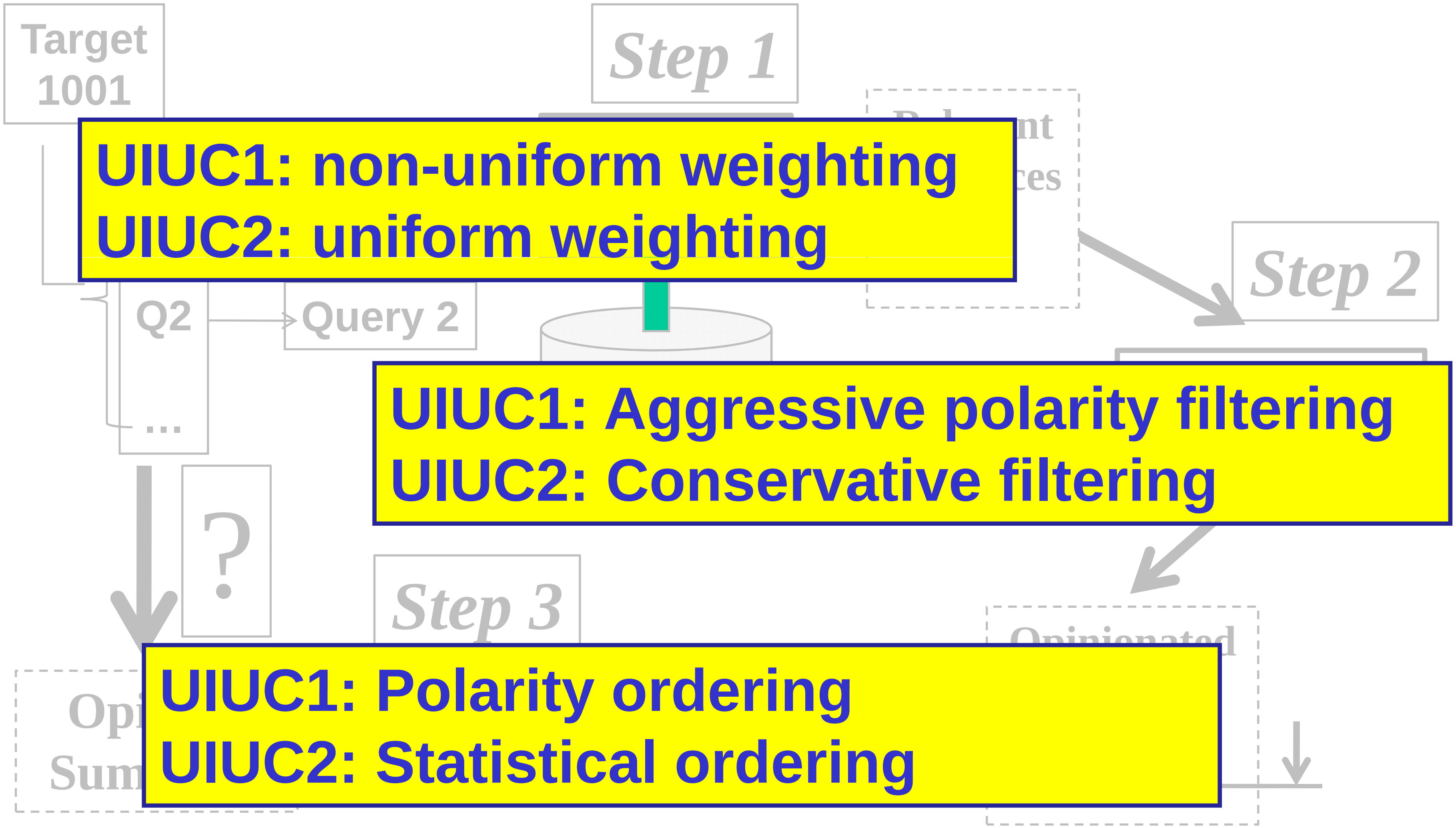


P(u,v) = (2+0.001)/(3*4+1.0)

General Approach



Submissions: UIUC1, UIUC2



UIUC1: non-uniform weighting
UIUC2: uniform weighting

UIUC1: Aggressive polarity filtering
UIUC2: Conservative filtering

UIUC1: Polarity ordering
UIUC2: Statistical ordering

Evaluation

- **Rank among runs without answer-snippet**

(Total: 19 runs)

	F-Score	Grammaticality	Non-redundancy	Structure/ Coherence	Fluency/ Readability	Responsiveness
UIUC1 Polarity	6	15	15	**5**	6	8
UIUC2 Coherence	**3**	9	10	15	16	**4**

Evaluation

- R … without

NE/NP retrieval, Polarity filtering

Polarity ordering

(Total: 19 runs)

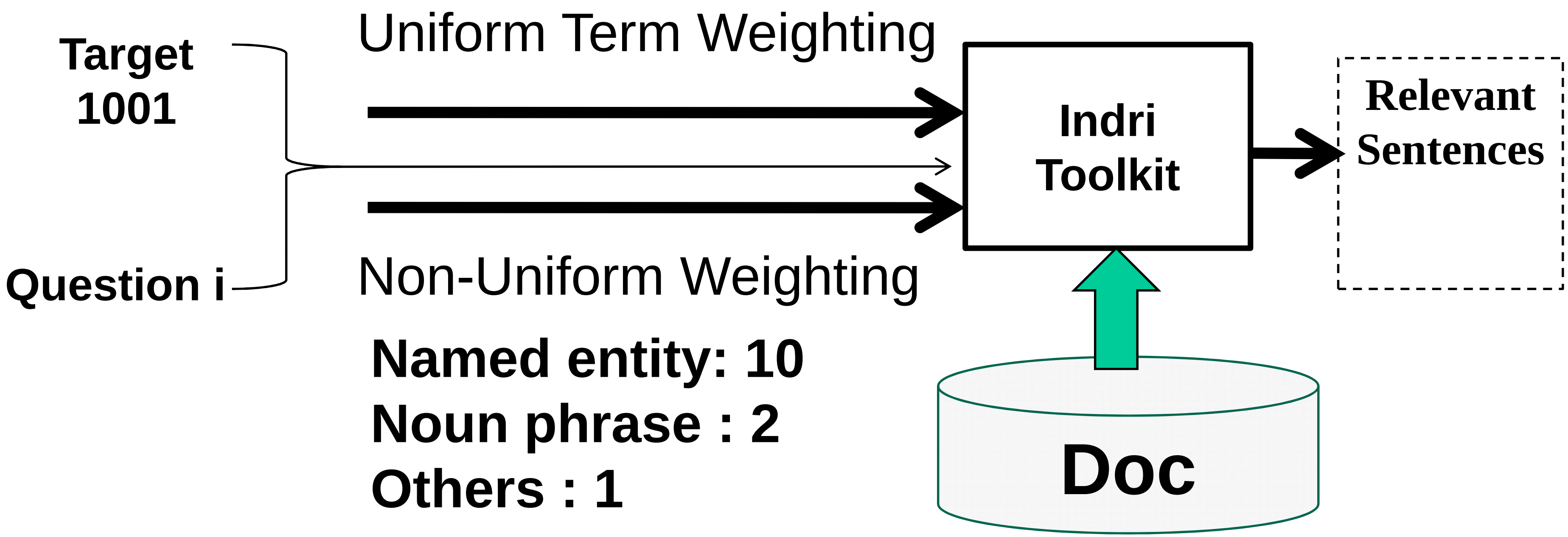
	…core	Grammaticality	Non-redundancy	Stru…e/ Cohe…nce	Fluency/ Readability	Responsiveness
UIUC1 Polarity	6	15	15	**5**	6	8
UIUC2 Coherence	**3**	9	10	15	16	**4**

Nothing

Statistical ordering

Evaluation of Named Entity Weighting

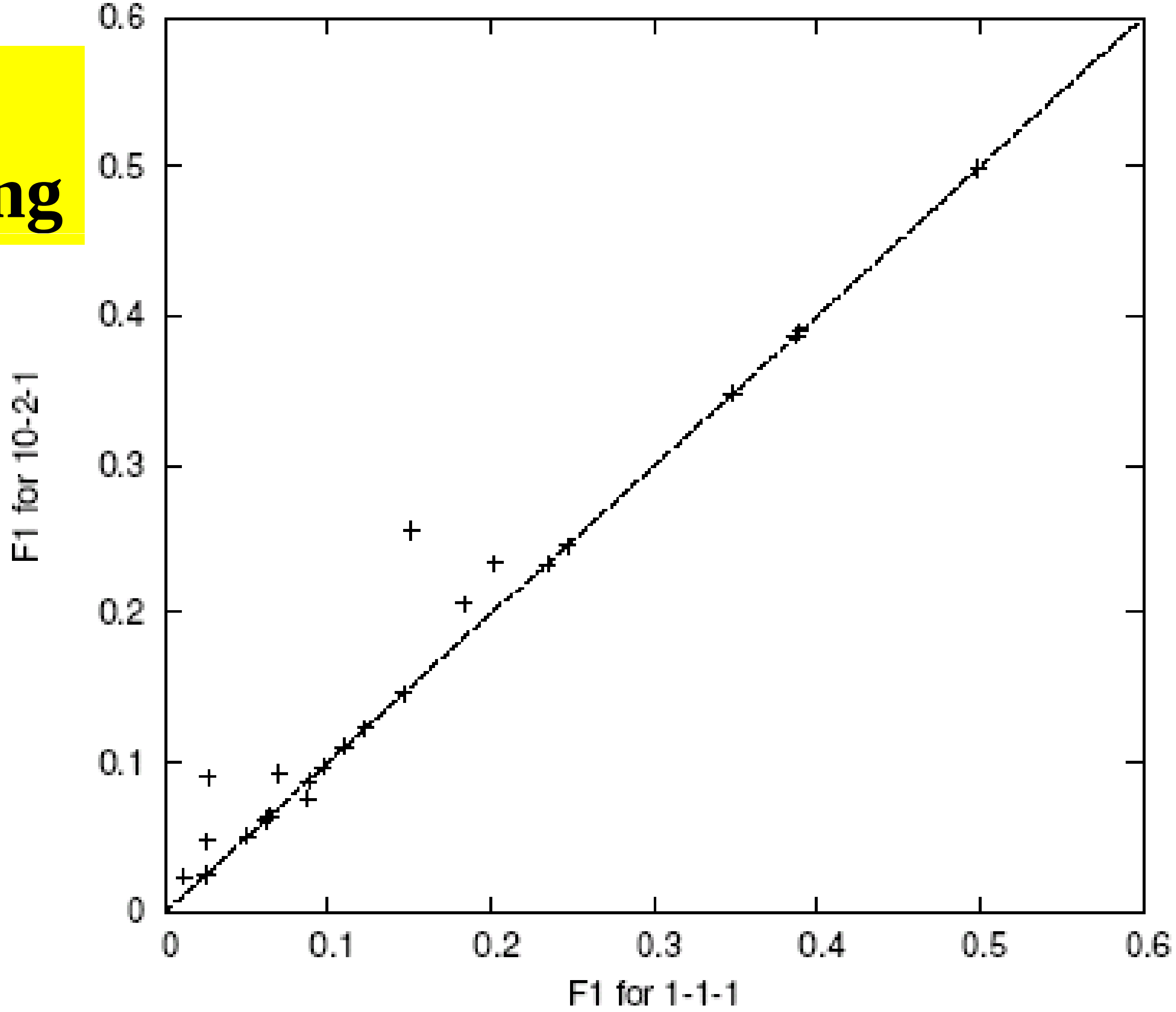
Assume a sentence is relevant iff
similarity(sentence, nugget description) > threshold



The Database and Information Systems Laboratory

Effectiveness of Entity Weighting

10-2-1 Weighting



1-1-1 Weighting

Polarity Module

- **Polarity module performance evaluation on the sentiment corpus. [Hu&Liu 04, Hu&Liu 04b]**

(Unit: # of sentence)

Classification result	Positive	Negative
NonOpinionated	1063	598
Positive	1363	371
Negative	383	412
Mixed	296	210
Total	**3105**	**1591**
Exact Match	**1363/3105=0.44**	**412/1591=0.26**
	(1363+412)/(3105+1591)=0.38	
Exact Opposite	**383/3105=0.12**	**371/1591=0.23**
	(383+371)/(3105+1591)=0.16	

Coherence optimization

- **Evaluation methods**
  - **Basic assumption**
    - **the sentence order of original document is coherent**
  - **Among given target documents,**
    use 70% as training set, 30% as test set.
  - **Measurement: strict pair matching**
    - **# of correct sentence pair / # of total adjacent sentence pair**

Probabilistic Coherence Function

$$c(s_i, s_j) = \sum_{u \in s_i, v \in s_j} \frac{p(u,v)}{|s_i||s_j|}$$

Average coherence probability over all word combinations

Point-wise Mutual information with smoothing

$$\hat{p}(u,v) = \frac{count("u\ and\ v\ in\ two\ adjacent\ sentences") + 0.001}{count(u) \cdot count(v) + 1.0}$$

Strict joint probability

$$\hat{p}(u,v) = \frac{count("u\ and\ v\ in\ two\ adjacent\ sentences") + 1}{"total\ count\ of\ word\ pairs" + (dictionary\ size)^2}$$

Probabilistic Coherence Function

Mutual information

$$\hat{p}(u,v) = p(u,v)\log(\frac{p(u,v)}{p(u)\,p(v)}) + p(not\,u,v)\log(\frac{p(not\,u,v)}{p(not\,u)\,p(v)})$$

$$+\, p(u,not\,v)\log(\frac{p(u,not\,v)}{p(u)\,p(not\,v)}) + p(not\,u,not\,v)\log(\frac{p(not\,u,not\,v)}{p(not\,u)\,p(not\,v)})$$

where,

$$p(u,v) = \frac{c(u,v)+0.25}{N+1},\ p(not\,u,v) = \frac{c(not\,u,v)+0.25}{N+1},$$

$$p(not\,u,v) = \frac{c(u,not\,v)+0.25}{N+1},\ p(not\,u,not\,v) = \frac{c(not\,u,not\,v)+0.25}{N+1},$$

N = c(u,v)+c(not u, v)+c(u, not v)+c(not u, not v)

For unseen pairs, p(u,v)=0.5*MIN(seen pairs in training)

Coherence optimization test

Selection of training words	Strict Joint Probability	Mutual Information	Pointwise Mutual Information
No Omission	0.022259	0.041651	**0.056063**
Omitted stopwords	**0.031389**	**0.054554**	**0.057119**
Omitted frequent words (counts > 33)	**0.031389**	**0.051460**	**0.049498**
(counts > 11)	0.027013	0.045725	0.046103
(counts > 6)	0.020448	0.032219	0.034785
(counts > 2)	0.019769	0.022259	0.021882
Omitted rare words (counts < 33)	0.022259	0.032898	0.044065
(counts < 14)	0.022033	0.032823	0.049045
(counts < 6)	0.021203	0.037048	0.054931
(counts < 2)	**0.022259**	**0.041802**	**0.056591**

– **Pointwise mutual information effectively penalize common words**

Coherence optimization test

- **Top ranked p(u,v) of strict joint probability**

u	v	p(u,v)
the	the	1.75E-003
to	the	1.22E-003
the	to	1.21E-003
the	of	1.14E-003
the	and	1.14E-003
of	the	1.13E-003
and	the	1.12E-003
a	the	1.06E-003
the	a	1.06E-003
…	…	…

  - **A lot of stopwords are top-ranked.**

Coherence optimization test

Selection of training words	Coherence score
Baseline: random order	0.01586
Strict Joint Probability	0.04308
Mutual Information	0.041651
Pointwise Mutual Information (UIUC2)	**0.056063**
Omitted stopwords	**0.057119**
Omitted non-stopwords	0.020750
Omitted 95% least frequent words (counts < 33):	0.044065
Omitted 90% least frequent words (counts < 14):	0.049045
Omitted 80% least frequent words (counts < 6):	0.054931
Omitted 60% least frequent words (counts < 2):	**0.056591**

- **Pointwise Mutual information was better than joint probability and normal mutual information.**
- **Eliminating common words, very rare words improved performance**

Conclusions

- **Limited improvement in retrieval performance using named entity and noun phrase**
- **Need for a good polarity classification module**
- **Possibility on the improvement of statistical sentence ordering module with different coherence function and word selection**

Thank you

University of Illinois at Urbana-Champaign
Hyun Duk Kim (hkim277@uiuc.edu)